ChatGPT 100,000 Patient 24-Month *In Silico* Phase III 5-Arm Pancreatic Cancer Clinical Trial Triplicate

Kevin Kawchak
Chief Executive Officer
ChemicalQDevice
San Diego, CA
July 24, 2025
kevink@chemicalqdevice.com

43.S57.REP.01.P43

Prompt 43

Gemini 2.5 Pro: 28 Pages, July 12, 2025

Settings: Temp=1, Thinking mode=On, Set thinking budget 32768, All Settings Off = (Structured output, Code execution, Function calling, Grounding with Google Search, URL context), All Settings Off=Safety settings, Output length=65536, Top P=0.95
g25p Tokens Input = 147,531, Tokens All = 154,075, Tokens Output = 6,544

Prompt 43 (Abbreviated):

"Instructions Start"
Analyze, utilize, and cite the provided documents to produce a comprehensive virtual study overview of the completed 100,000-patient virtual triplicate simulations. Produce a single, detailed report in the "Executive Summary", "Technical Details", "Key Insights" format. Use large, interpretable markdown tables designated with appropriate rows R1, R2.. and columns C1, C2.. suitable for downstream data extraction and visualization.

Input Files for Processing:

- Trial reports, log file verifications, external validations, and visualizations for a 100,000-patient, 5-arm *in-silico* Phase III simulation run in triplicate and verified by multiple AI models (grk4, grk3, ops4, g25p, o3pr). Log file verfications correspond to files such as S35.VER.02.P32.

A. Virtual Study Triplicate Details

1. **Four Tables with specific rows R1, R2.. and columns C1, C2..**
2. **Fill in Details of each cell with combined data from the included files below**
3. **Table 01: 3 Virtual Trials - Provide Additional Details**

- **Study Title/Identifier**
- **Primary Goal**
- **Trial Phase Equivalence** (Phase III details)
- **Study Design** (5-arm in-silico simulation)
- **Trial Arms** (List the specific arms for each)
- **Patient Population Size**
- **Patient Archetypes** (7 archetypes)

4. **Table 02: 3 Virtual Trial Details - Provide Additional Details**
    - **Drug Combination(s)** (Note the shared core triplet)
    - **Patient Data Granularity** (Describe the level of detail for virtual patient creation)
    - **Modeling Architecture** (100K trial's exponential survival model)
    - **Project Timeline**
    - **Primary Endpoints**
    - **Key AI Models Utilized** (List for both, based on the provided information)
5. **Table 03: Benefits and Drawbackss - Provide Additional Details**
    - **Itemized Benefits** of the Completed 100K Patient Triplicate Simulation:
      Pay particular attention to all benefits derived from the completed triplicate simulation. Analyze the value of its speed, scale, and robust cross-model/cross-trial verification (as seen in files such as S43, S48, S49, S50, S55, S56).
    - **Itemized Drawbacks** of the Completed 100K Patient Triplicate Simulation:
      List the drawbacks and limitations of the 100K trial's approach, considering factors like its simplified patient models and potential for "black-box" objections. Detail how methods used for the simulated trials could be improved in future studies.
6. **Table 04: Reproducibility Findings - Most Comprehensively Detailed Table**
    - **Validation** (100K trial's internal log verification and external validation against Flatiron data) Provide full detail regarding all results reported. Be sure to include inclusion of data synergies from Table T1 - OS concordance, Table T2 - OS Summary Metrics, and Table T3-ECOG Confidence scores.
    - **Reproducibility** (Overall reproducibility metrics of triplicate runs and cross-model verification in files like S43, S50, S55, S56). Include analysis of visualization scripts in Python from files like S48.VIS.01.P38. Be sure to include in full detail how reproducibility across the three simulated trials was observed or not observed by analyzing, utilizing, and citing specific documents included below.

“Instructions End”
“Use Model Abbreviations in Output Start”
grk4 = Grok 4, grk3 = Grok 3, ops4 = Opus 4, g25p = Gemini 2.5 Pro, o3pr = ChatGPT o3-pro, o3ph = ChatGPT o3-pro Research
“Use Model Abbreviations in Output End”

”File Descriptions Start”
S33.TRL.13.P30 = Trial 1 (Example)
S35.VER.02.P32 = Trial 1 tables vs. log file verifications (Example)
S35b.VER.03.P34 = Trial 1 external validation of log file (Example)
S36.VIS.01.P33 = Trial 1 visualizations (Example)
S37.TRL.14.P30 = Trial 2
S38.VER.01.P32 = Trial 2 tables vs. log file verifications
S38b.VER.02.P35 = Trial 2 external validation of log file

S39.VIS.01.P33 = Trial 2 visualizations
S40.TRL.15.P30 = Trial 3
S41.VER.01.P32 = Trial 3 tables vs. log file verifications
S41b.VER.02.P36 = Trial 3 external validation of log file
S42.VIS.01.P33 = Trial 3 visualizations
S43.TST.01.P37 = grk4 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output) (Example)
S44.TST.02.P37 = grk3 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S45.TST.03.P37 = ops4 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S46.TST.04.P37 = g25p 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S47.TST.05.P37 = o3pr 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S48.VIS.01.P38 = 3 Trial Tables Cross-Verification: Visualize Models (Example)
S49.VIS.02.P39 = 3 Trial Tables Cross-Verification: Visualize Trials (Example)
S50.TST.01.P40 = grk4 Meta-Verification Tables Cross-Trial of Dataset 3 S35.VER.02.P32, S38.VER.01.P32,
S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output) (Example)
S51.TST.02.P40 = grk3 Meta-Verification Tables Cross-Trial of Dataset 3 S35.VER.02.P32, S38.VER.01.P32,
S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S52.TST.03.P40 = ops4 Meta-Verification Tables Cross-Trial of Dataset 3 S35.VER.02.P32, S38.VER.01.P32,
S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S53.TST.04.P40 = g25p Meta-Verification Tables Cross-Trial of Dataset 3 S35.VER.02.P32, S38.VER.01.P32,
S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S54.TST.05.P40 = o3pr Meta-Verification Tables Cross-Trial of Dataset 3 S35.VER.02.P32, S38.VER.01.P32,
S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S55.VIS.01.P41 = Verification Tables Cross-Trial: Visualize Models (Example)
S56.VIS.02.P42 = Verification Tables Cross-Trial: Visualize Trials (Example)
”File Descriptions End”

[”Files Start”]

User: Find/Remove

Generation 57 (Formatted)

Gemini 2.5 Pro

Virtual Study Overview: Analysis of 100,000-Patient Triplicate Simulations

---

Executive Summary

This report provides a comprehensive overview of a completed virtual study involving triplicate simulations of a 100,000-patient, 5-arm, Phase III-equivalent clinical trial in advanced Pancreatic Ductal Adenocarcinoma (PDAC). The study evaluated a novel triplet therapy (Daraxonrasib + Mitazalimab + liposomal Irinotecan) against various doublet combinations and a standard-of-care control. Each of the three simulation runs was independently verified against its source logs and validated against external real-world data. The entire process was cross-verified by a suite of five independent AI models (grk4, grk3, ops4, g25p, o3pr) to ensure maximum robustness and reproducibility.

The clinical findings from the simulation consistently demonstrated that the triplet therapy (Arm A) conferred the greatest efficacy, with a mean Overall Survival (OS) of approximately 8.7 months versus 6.1 months for the control arm (OS Hazard Ratio [HR] ~0.69). This significant survival benefit was, however, accompanied by the highest rate of severe (Grade ≥3) adverse events (94% vs. 76.5% in control), highlighting a critical efficacy-toxicity trade-off.

Methodologically, the study establishes a new benchmark for in-silico trial validation. The triplicate runs were found to be exceptionally reproducible, with cross-trial consistency scores for key metrics consistently rated above 8.5 out of 10 by all AI models (S43-S47). External validation of the control arm against Flatiron data showed high concordance for survival outcomes (Pearson r = 0.999) but revealed a mismatch in baseline ECOG performance status, indicating an area for future model refinement (S35b, S38b, S41b). The multi-AI model verification process itself was highly reliable, confirming the stability of the simulation and analysis pipeline and underscoring the value of this approach for generating robust, defensible virtual evidence.

Technical Details

Table 01: 3 Virtual Trials - Overview

C1: Study Title/Identifier	C2: Primary Goal	C3: Trial Phase Equivalence	C4: Study Design	C5: Trial Arms	C6: Patient Population Size	C7: Patient Archetypes

R1: Details	A Phase III Virtual Study of Triplet Daraxonrasib + Mitazalimab + liposomal Irinotecan vs Doublets vs Chemotherapy in Advanced Pancreatic Ductal Adenocarcinoma (PDAC-SIM-001)	To compare the efficacy and safety of a novel triplet therapy against doublet combinations and standard chemotherapy control in advanced PDAC.	**Phase:** III (Virtual Simulation) **Design:** Randomized, controlled, parallel-group, five-arm study. **Endpoints:** Co-primary endpoints of Overall Survival (OS) and Progression-Free Survival (PFS) with a 24-month data cutoff.	5-arm in-silico simulation based on predefined patient archetypes and time-to-event models. Patients were randomized 1:1:1:1:1.	**Arm A:** Triplet (Daraxonrasib + Mitazalimab + liposomal Irinotecan) **Arm B:** Doublet (Mitazalimab + liposomal Irinotecan) **Arm C:** Doublet (Daraxonrasib + liposomal Irinotecan) **Arm D:** Doublet (Daraxonrasib + Mitazalimab) **Arm E:** Control (nal-IRI + 5-FU chemotherapy)	**Total:** 100,000 virtual patients per simulation run, conducted in triplicate. **Per Arm:** 20,000 patients.	**7 Predefined Archetypes:** ARCH-01: Young_Fit_Metastatic ARCH-02: Elderly_Frail_Metastatic ARCH-03: LAPC_Standard_Fitness ARCH-04: Young_Fit_BRCAm ARCH-05: Metastatic_KRAS_G12C ARCH-06: Metastatic_High_Stroma ARCH-07: Advanced_Refractory_PS1

Source: Synthesized from trial reports S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30.

Table 02: 3 Virtual Trials - Technical Specifications

C1: Drug Combination(s)	C2: Patient Data Granularity	C3: Modeling Architecture	C4: Project Timeline	C5: Primary Endpoints	C6: Key AI Models Utilized

R1: Details	**Core Triplet:** Daraxonrasib (KRAS G12C inhibitor) + Mitazalimab (immunotherapy) + liposomal Irinotecan. **Doublets & Control:** Various combinations of the core agents and a standard chemotherapy control were tested across the 5 arms.	Virtual patients were generated with a rich set of features defined by seven archetypes. Key data points included: age, disease stage (metastatic vs. locally advanced), ECOG performance status (0, 1, 2), tumor genomics (KRAS mutation status, specifically G12C; germline BRCA mutation status), and baseline tumor markers (CA 19-9).	An exponential survival model (Weibull shape k=1.0) was used to simulate time-to-event outcomes. Baseline hazards for the control arm were set to achieve median PFS of 3.1 months and OS of 6.1 months. Multiplicative hazard ratios (HRs) for each drug and a synergy factor (0.90) for the triplet were applied to model treatment effects.	The virtual trial simulations and analyses were conducted with a report date of July-August 2025. A fixed random seed (20250624) was used across all three trials to ensure reproducibility of the simulation runs.	**Co-primary Endpoints:** 1. **Overall Survival (OS):** Time from randomization to death from any cause. 2. **Progression-Free Survival (PFS):** Time from randomization to disease progression or death. **Secondary Endpoints:** 12-month OS rates and incidence of Grade ≥3 adverse events.	**For Cross-Verification & Meta-Verification:** 1. **grk4:** Grok 4 2. **grk3:** Grok 3 3. **ops4:** Opus 4 4. **g25p:** Gemini 2.5 Pro 5. **o3pr:** ChatGPT o3-pro

Source: Synthesized from trial reports S33, S37, S40 and verification files S43-S56.

Table 03: Benefits and Drawbacks of the 100K Patient Triplicate Simulation

C1: Itemized Benefits	C2: Itemized Drawbacks & Limitations

R1: Speed & Scale	**Unprecedented Scale:** Enabled the analysis of 100,000 patients per run, a scale infeasible for traditional trials, providing highly stable estimates for medians and HRs. **Rapid Execution:** The entire triplicate simulation and verification process was completed in a fraction of the time required for a physical trial, accelerating hypothesis testing.	**Model Simplification:** The exponential survival model is a simplification of complex real-world patient trajectories. It did not account for factors like dose modifications, treatment discontinuations due to AEs, or competing risks (as noted in S33).
R2: Robustness & Verification	**Triplicate Runs:** Running the simulation three times established the stability and low variance of the simulation engine. Cross-trial consistency scores were consistently high (S43-S47). **Multi-Model AI Verification:** The use of five diverse AI models (grk4, grk3, ops4, g25p, o3pr) to cross-verify all data provided an exceptionally robust, multi-perspective validation of the results and the verification process itself (S48, S55).	**Patient Profile Mismatch:** External validation against Flatiron data revealed a significant mismatch in baseline ECOG distribution, with the simulation having a healthier patient profile (fewer ECOG 2 patients). This limits the direct generalizability of the results to a real-world, less-fit population (S35b, S38b, S41b).
R3: Hypothesis Testing & Subgroup Analysis	**Complex Design Testing:** Allowed for the simultaneous evaluation of five complex regimens (a triplet and three doublets vs. control), which would be logistically challenging and costly in a traditional trial. **Deep Subgroup Analysis:** The large sample size enabled statistically robust analysis of seven distinct patient archetypes, confirming treatment benefits across subgroups and identifying a key treatment-genotype interaction (ARCH-05 / KRAS G12C).	**Reporting & Data Discrepancies:** The verification process consistently identified a major discrepancy between the reported KRAS-mutant (%) in the CSRs (~91%) and the kras_g12c=1 flag in the logs (~5%), indicating a need for clearer variable definitions and data capture protocols (S35, S38, S41).

R4: Future Improvements	**Data-Driven Trial Design:** The findings provide a quantitative framework to inform the design of future real-world trials, helping to prioritize the most promising regimens (e.g., Arm A vs. Arm D) and patient selection strategies.	**Refinement Opportunities:** Future simulations should incorporate more sophisticated models (e.g., agent-based models), include dynamic responses to toxicity, model biomarker efficacy more granularly (e.g., KRAS G12C benefit only applying to that subgroup), and refine the patient generation process to better match real-world ECOG distributions.

Source: Analysis of all provided files, with specific drawbacks cited.

Table 04: Reproducibility and Validation Findings

	C1: Validation (External Concordance)	C2: Reproducibility (Internal & Cross-Model Consistency)
R1: Overall Survival (OS)	**High Concordance:** The control arm (Arm E) from all three simulations demonstrated high external validity against Flatiron real-world data. OS% at all measured time points (0-24 months) fell within the ±5% pre-specified concordance threshold. The mean OS% difference was ~1.5% and the Pearson correlation was 0.999, both passing validation criteria (S35b, S38b, S41b).	**High Reproducibility:** Median OS values were extremely stable across the triplicate runs (e.g., Arm A mean OS of 8.73 mo, with a range of only 0.1 mo). Cross-trial consistency scores for Median OS and OS HR were high, averaging 8.98 and 9.08 respectively across the five AI models (S43-S47). This indicates the OS outcomes were highly reproducible.

R2: Baseline Characteristics	**Partial Concordance:** The simulated ECOG performance status distribution failed external validation. The absolute differences for ECOG 0, 1, and 2 vs. Flatiron data were ~5%, ~14%, and ~19% respectively, all exceeding the ±5% failure threshold (S35b, S38b, S41b). This indicates the simulated patient population was fitter than the real-world cohort.	**Exceptional Reproducibility:** Baseline characteristics were nearly identical across the three trials. Cross-trial consistency scores for all baseline metrics were ≥9.8 out of 10 across all AI models (S43-S47). The meta-verification analysis of the verification logs confirmed that discrepancies found were also highly consistent; for example, the KRAS-mutant deviation was found with a consistency score of 10.0 (S50-S54).
R3: Cross-Model Verification & Analysis	Not Applicable. External validation was performed on the simulation output itself, not on the AI models' analysis.	**Strong Inter-Model Agreement:** The five AI models showed remarkable agreement in their analyses. Visualizations confirmed a "tight cluster" for grk4, g25p, and o3pr, with grk3 and ops4 as minor outliers (S55). Agreement was highest for baseline metrics and lowest for archetype-specific outcomes (S48). The analysis included programmatic generation of visualizations (e.g., 01_heatmap_consistency_scores.py from S48) to quantify this agreement.
R4: Overall Reproducibility Assessment	The simulation's survival dynamics are externally valid, but the patient profile has limitations.	**Highly Robust:** The triplicate runs were highly consistent, with minimal variance in all primary and secondary endpoints. The AI-driven cross-trial verification process confirmed this stability with high consistency scores. Furthermore, meta-verification of the verification logs themselves also scored highly (mean scores >8.8), confirming the entire data generation and analysis pipeline is robust and reproducible (S50-S56). Analysis of visualization scripts (S49, S56) showed that percentage-based metrics (like AE rates) had higher consistency than time-to-event metrics (like median OS).

Source: Synthesis of all verification (S35, S38, S41), external validation (S35b, S38b, S41b), cross-trial verification (S43-S47, S48, S49), and meta-verification (S50-S54, S55, S56) files.

Key Insights

1. **Clinical Strategy:** The simulations provide strong evidence that a multi-pronged therapeutic strategy combining targeted therapy, immunotherapy, and chemotherapy (Arm A) yields the most significant survival benefit in advanced PDAC. The OS improvement of ~2.6 months over control is clinically meaningful. However, the associated 94% incidence of severe AEs makes this a high-risk, high-reward strategy. The chemo-free doublet of Daraxonrasib + Mitazalimab (Arm D) emerges as a highly promising alternative, offering a substantial survival benefit (OS HR ~0.76) with a more manageable toxicity profile, making it a potentially superior option for less-fit patients.
2. **Biomarker Importance:** The archetype sub-analyses powerfully validated the importance of biomarker-driven therapy. Patients in the KRAS G12C subgroup (ARCH-05) derived a clear and substantial benefit only from regimens containing the targeted inhibitor Daraxonrasib. This finding underscores the necessity of molecular profiling in PDAC to match patients with the most effective treatments and avoid ineffective therapies.
3. **Methodological Robustness:** This study exemplifies a new standard in virtual trial validation. By running the simulation in triplicate, performing internal log verification, validating against external real-world data, and employing a panel of five diverse AI models for cross-verification, the project established a high degree of confidence in the results. The process was robust enough to not only confirm the stability of the outputs but also to consistently identify its own limitations, such as the discrepancy in the KRAS definition and the mismatch in the ECOG profile.
4. **AI in Clinical Analysis:** The multi-model AI verification demonstrated both the power and the nuances of using AI for clinical data analysis. While there was remarkable consensus on objective metrics, minor systematic differences were observed between models, particularly in the calculation of variance-based consistency scores (S55, S56). This highlights the importance of using a multi-model approach or standardized consensus methods for critical analyses, mitigating the risk of relying on a single "black-box" algorithm. The success of this verification-of-verification process pioneers a new level of rigor for in-silico evidence generation.

User: Find/Remove

Generation 57 (Plain Text)

Gemini 2.5 Pro

Virtual Study Overview: Analysis of 100,000-Patient Triplicate Simulations
Executive Summary

This report provides a comprehensive overview of a completed virtual study involving triplicate simulations of a 100,000-patient, 5-arm, Phase III-equivalent clinical trial in advanced Pancreatic Ductal Adenocarcinoma (PDAC). The

study evaluated a novel triplet therapy (Daraxonrasib + Mitazalimab + liposomal Irinotecan) against various doublet combinations and a standard-of-care control. Each of the three simulation runs was independently verified against its source logs and validated against external real-world data. The entire process was cross-verified by a suite of five independent AI models (grk4, grk3, ops4, g25p, o3pr) to ensure maximum robustness and reproducibility.

The clinical findings from the simulation consistently demonstrated that the triplet therapy (Arm A) conferred the greatest efficacy, with a mean Overall Survival (OS) of approximately 8.7 months versus 6.1 months for the control arm (OS Hazard Ratio [HR] ~0.69). This significant survival benefit was, however, accompanied by the highest rate of severe (Grade ≥3) adverse events (94% vs. 76.5% in control), highlighting a critical efficacy-toxicity trade-off.

Methodologically, the study establishes a new benchmark for in-silico trial validation. The triplicate runs were found to be exceptionally reproducible, with cross-trial consistency scores for key metrics consistently rated above 8.5 out of 10 by all AI models (S43-S47). External validation of the control arm against Flatiron data showed high concordance for survival outcomes (Pearson r = 0.999) but revealed a mismatch in baseline ECOG performance status, indicating an area for future model refinement (S35b, S38b, S41b). The multi-AI model verification process itself was highly reliable, confirming the stability of the simulation and analysis pipeline and underscoring the value of this approach for generating robust, defensible virtual evidence.

Technical Details

Table 01: 3 Virtual Trials - Overview

C1: Study Title/Identifier C2: Primary Goal C3: Trial Phase Equivalence C4: Study Design C5: Trial Arms C6: Patient Population Size C7: Patient Archetypes

R1: Details A Phase III Virtual Study of Triplet Daraxonrasib + Mitazalimab + liposomal Irinotecan vs Doublets vs Chemotherapy in Advanced Pancreatic Ductal Adenocarcinoma (PDAC-SIM-001) To compare the efficacy and safety of a novel triplet therapy against doublet combinations and standard chemotherapy control in advanced PDAC. Phase: III (Virtual Simulation) Design: Randomized, controlled, parallel-group, five-arm study. Endpoints: Co-primary endpoints of Overall Survival (OS) and Progression-Free Survival (PFS) with a 24-month data cutoff. 5-arm in-silico simulation based on predefined patient archetypes and time-to-event models. Patients were randomized 1:1:1:1:1. Arm A: Triplet (Daraxonrasib + Mitazalimab + liposomal Irinotecan) Arm B: Doublet (Mitazalimab + liposomal Irinotecan) Arm C: Doublet (Daraxonrasib + liposomal Irinotecan) Arm D: Doublet (Daraxonrasib + Mitazalimab) Arm E: Control (nal-IRI + 5-FU chemotherapy) Total: 100,000 virtual patients per simulation run, conducted in triplicate. Per Arm: 20,000 patients.7 Predefined Archetypes: ARCH-01: Young_Fit_Metastatic ARCH-02: Elderly_Frail_Metastatic ARCH-03: LAPC_Standard_Fitness ARCH-04: Young_Fit_BRCAm ARCH-05: Metastatic_KRAS_G12C ARCH-06: Metastatic_High_Stroma ARCH-07: Advanced_Refractory_PS1

Source: Synthesized from trial reports S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30.

Table 02: 3 Virtual Trials - Technical Specifications

C1: Drug Combination(s) C2: Patient Data Granularity C3: Modeling Architecture C4: Project Timeline C5: Primary Endpoints C6: Key AI Models Utilized

R1: Details Core Triplet: Daraxonrasib (KRAS G12C inhibitor) + Mitazalimab (immunotherapy) + liposomal Irinotecan. Doublets & Control: Various combinations of the core agents and a standard chemotherapy control were tested across the 5 arms. Virtual patients were generated with a rich set of features defined by seven archetypes. Key data points included: age, disease stage (metastatic vs. locally advanced), ECOG performance status (0, 1, 2), tumor genomics (KRAS mutation status, specifically G12C; germline BRCA mutation status), and baseline tumor markers (CA 19-9). An exponential survival model (Weibull shape k=1.0) was used to simulate time-to-event outcomes. Baseline hazards for the control arm were set to achieve median PFS of 3.1 months and OS of 6.1 months. Multiplicative hazard ratios (HRs) for each drug and a synergy factor (0.90) for the triplet were applied to model treatment effects. The virtual trial

simulations and analyses were conducted with a report date of July-August 2025. A fixed random seed (20250624) was used across all three trials to ensure reproducibility of the simulation runs. Co-primary Endpoints: 1. Overall Survival (OS): Time from randomization to death from any cause. 2. Progression-Free Survival (PFS): Time from randomization to disease progression or death. Secondary Endpoints: 12-month OS rates and incidence of Grade ≥3 adverse events. For Cross-Verification & Meta-Verification: 1. grk4: Grok 4 2. grk3: Grok 3 3. ops4: Opus 4 4. g25p: Gemini 2.5 Pro 5. o3pr: ChatGPT o3-pro

Source: Synthesized from trial reports S33, S37, S40 and verification files S43-S56.

Table 03: Benefits and Drawbacks of the 100K Patient Triplicate Simulation

C1: Itemized Benefits C2: Itemized Drawbacks & Limitations

R1: Speed & Scale Unprecedented Scale: Enabled the analysis of 100,000 patients per run, a scale infeasible for traditional trials, providing highly stable estimates for medians and HRs. Rapid Execution: The entire triplicate simulation and verification process was completed in a fraction of the time required for a physical trial, accelerating hypothesis testing. Model Simplification: The exponential survival model is a simplification of complex real-world patient trajectories. It did not account for factors like dose modifications, treatment discontinuations due to AEs, or competing risks (as noted in S33).

R2: Robustness & Verification Triplicate Runs: Running the simulation three times established the stability and low variance of the simulation engine. Cross-trial consistency scores were consistently high (S43-S47). Multi-Model AI Verification: The use of five diverse AI models (grk4, grk3, ops4, g25p, o3pr) to cross-verify all data provided an exceptionally robust, multi-perspective validation of the results and the verification process itself (S48, S55). Patient Profile Mismatch: External validation against Flatiron data revealed a significant mismatch in baseline ECOG distribution, with the simulation having a healthier patient profile (fewer ECOG 2 patients). This limits the direct generalizability of the results to a real-world, less-fit population (S35b, S38b, S41b).

R3: Hypothesis Testing & Subgroup Analysis Complex Design Testing: Allowed for the simultaneous evaluation of five complex regimens (a triplet and three doublets vs. control), which would be logistically challenging and costly in a traditional trial. Deep Subgroup Analysis: The large sample size enabled statistically robust analysis of seven distinct patient archetypes, confirming treatment benefits across subgroups and identifying a key treatment-genotype interaction (ARCH-05 / KRAS G12C). Reporting & Data Discrepancies: The verification process consistently identified a major discrepancy between the reported KRAS-mutant (%) in the CSRs (~91%) and the kras_g12c=1 flag in the logs (~5%), indicating a need for clearer variable definitions and data capture protocols (S35, S38, S41).

R4: Future Improvements Data-Driven Trial Design: The findings provide a quantitative framework to inform the design of future real-world trials, helping to prioritize the most promising regimens (e.g., Arm A vs. Arm D) and patient selection strategies. Refinement Opportunities: Future simulations should incorporate more sophisticated models (e.g., agent-based models), include dynamic responses to toxicity, model biomarker efficacy more granularly (e.g., KRAS G12C benefit only applying to that subgroup), and refine the patient generation process to better match real-world ECOG distributions.

Source: Analysis of all provided files, with specific drawbacks cited.

Table 04: Reproducibility and Validation Findings

C1: Validation (External Concordance) C2: Reproducibility (Internal & Cross-Model Consistency)

R1: Overall Survival (OS) High Concordance: The control arm (Arm E) from all three simulations demonstrated high external validity against Flatiron real-world data. OS% at all measured time points (0-24 months) fell within the ±5% pre-specified concordance threshold. The mean OS% difference was ~1.5% and the Pearson correlation was 0.999, both passing validation criteria (S35b, S38b, S41b). High Reproducibility: Median OS values were extremely stable across the triplicate runs (e.g., Arm A mean OS of 8.73 mo, with a range of only 0.1 mo). Cross-trial consistency scores

for Median OS and OS HR were high, averaging 8.98 and 9.08 respectively across the five AI models (S43-S47). This indicates the OS outcomes were highly reproducible.
R2: Baseline Characteristics Partial Concordance: The simulated ECOG performance status distribution failed external validation. The absolute differences for ECOG 0, 1, and 2 vs. Flatiron data were ~5%, ~14%, and ~19% respectively, all exceeding the ±5% failure threshold (S35b, S38b, S41b). This indicates the simulated patient population was fitter than the real-world cohort. Exceptional Reproducibility: Baseline characteristics were nearly identical across the three trials. Cross-trial consistency scores for all baseline metrics were ≥9.8 out of 10 across all AI models (S43-S47). The meta-verification analysis of the verification logs confirmed that discrepancies found were also highly consistent; for example, the KRAS-mutant deviation was found with a consistency score of 10.0 (S50-S54).
R3: Cross-Model Verification & Analysis Not Applicable. External validation was performed on the simulation output itself, not on the AI models' analysis. Strong Inter-Model Agreement: The five AI models showed remarkable agreement in their analyses. Visualizations confirmed a "tight cluster" for grk4, g25p, and o3pr, with grk3 and ops4 as minor outliers (S55). Agreement was highest for baseline metrics and lowest for archetype-specific outcomes (S48). The analysis included programmatic generation of visualizations (e.g., 01_heatmap_consistency_scores.py from S48) to quantify this agreement.
R4: Overall Reproducibility Assessment The simulation's survival dynamics are externally valid, but the patient profile has limitations. Highly Robust: The triplicate runs were highly consistent, with minimal variance in all primary and secondary endpoints. The AI-driven cross-trial verification process confirmed this stability with high consistency scores. Furthermore, meta-verification of the verification logs themselves also scored highly (mean scores >8.8), confirming the entire data generation and analysis pipeline is robust and reproducible (S50-S56). Analysis of visualization scripts (S49, S56) showed that percentage-based metrics (like AE rates) had higher consistency than time-to-event metrics (like median OS).

Source: Synthesis of all verification (S35, S38, S41), external validation (S35b, S38b, S41b), cross-trial verification (S43-S47, S48, S49), and meta-verification (S50-S54, S55, S56) files.

Key Insights

Clinical Strategy: The simulations provide strong evidence that a multi-pronged therapeutic strategy combining targeted therapy, immunotherapy, and chemotherapy (Arm A) yields the most significant survival benefit in advanced PDAC. The OS improvement of ~2.6 months over control is clinically meaningful. However, the associated 94% incidence of severe AEs makes this a high-risk, high-reward strategy. The chemo-free doublet of Daraxonrasib + Mitazalimab (Arm D) emerges as a highly promising alternative, offering a substantial survival benefit (OS HR ~0.76) with a more manageable toxicity profile, making it a potentially superior option for less-fit patients.

Biomarker Importance: The archetype sub-analyses powerfully validated the importance of biomarker-driven therapy. Patients in the KRAS G12C subgroup (ARCH-05) derived a clear and substantial benefit only from regimens containing the targeted inhibitor Daraxonrasib. This finding underscores the necessity of molecular profiling in PDAC to match patients with the most effective treatments and avoid ineffective therapies.

Methodological Robustness: This study exemplifies a new standard in virtual trial validation. By running the simulation in triplicate, performing internal log verification, validating against external real-world data, and employing a panel of five diverse AI models for cross-verification, the project established a high degree of confidence in the results. The process was robust enough to not only confirm the stability of the outputs but also to consistently identify its own limitations, such as the discrepancy in the KRAS definition and the mismatch in the ECOG profile.

AI in Clinical Analysis: The multi-model AI verification demonstrated both the power and the nuances of using AI for clinical data analysis. While there was remarkable consensus on objective metrics, minor systematic differences were

observed between models, particularly in the calculation of variance-based consistency scores (S55, S56). This highlights the importance of using a multi-model approach or standardized consensus methods for critical analyses, mitigating the risk of relying on a single "black-box" algorithm. The success of this verification-of-verification process pioneers a new level of rigor for in-silico evidence generation.

ORIGINAL

Generation 57 (Formatted)

Gemini 2.5 Pro

1969 words, 1 minutes 18.95 seconds

Virtual Study Overview: Analysis of 100,000-Patient Triplicate Simulations

---

Executive Summary

This report provides a comprehensive overview of a completed virtual study involving triplicate simulations of a 100,000-patient, 5-arm, Phase III-equivalent clinical trial in advanced Pancreatic Ductal Adenocarcinoma (PDAC). The study evaluated a novel triplet therapy (Daraxonrasib + Mitazalimab + liposomal Irinotecan) against various doublet combinations and a standard-of-care control. Each of the three simulation runs was independently verified against its source logs and validated against external real-world data. The entire process was cross-verified by a suite of five independent AI models (grk4, grk3, ops4, g25p, o3pr) to ensure maximum robustness and reproducibility.

The clinical findings from the simulation consistently demonstrated that the triplet therapy (Arm A) conferred the greatest efficacy, with a mean Overall Survival (OS) of approximately 8.7 months versus 6.1 months for the control arm (OS Hazard Ratio [HR] ~0.69). This significant survival benefit was, however, accompanied by the highest rate of severe (Grade ≥3) adverse events (94% vs. 76.5% in control), highlighting a critical efficacy-toxicity trade-off.

Methodologically, the study establishes a new benchmark for in-silico trial validation. The triplicate runs were found to be exceptionally reproducible, with cross-trial consistency scores for key metrics consistently rated above 8.5 out of 10 by all AI models (S43-S47). External validation of the control arm against Flatiron data showed high concordance for survival outcomes (Pearson r = 0.999) but revealed a mismatch in baseline ECOG performance status, indicating an area for future model refinement (S35b, S38b, S41b). The multi-AI model verification process itself was highly reliable, confirming the stability of the simulation and analysis pipeline and underscoring the value of this approach for generating robust, defensible virtual evidence.

Technical Details

Table 01: 3 Virtual Trials - Overview

C1: Study Title/Identifier	C2: Primary Goal	C3: Trial Phase Equivalence	C4: Study Design	C5: Trial Arms	C6: Patient Population Size	C7: Patient Archetypes

R1: Details	A Phase III Virtual Study of Triplet Daraxonrasib + Mitazalimab + liposomal Irinotecan vs Doublets vs Chemotherapy in Advanced Pancreatic Ductal Adenocarcinoma (PDAC-SIM-001)	To compare the efficacy and safety of a novel triplet therapy against doublet combinations and standard chemotherapy control in advanced PDAC.	**Phase:** III (Virtual Simulation) **Design:** Randomized, controlled, parallel-group, five-arm study. **Endpoints:** Co-primary endpoints of Overall Survival (OS) and Progression-Free Survival (PFS) with a 24-month data cutoff.	5-arm in-silico simulation based on predefined patient archetypes and time-to-event models. Patients were randomized 1:1:1:1:1.	**Arm A:** Triplet (Daraxonrasib + Mitazalimab + liposomal Irinotecan) **Arm B:** Doublet (Mitazalimab + liposomal Irinotecan) **Arm C:** Doublet (Daraxonrasib + liposomal Irinotecan) **Arm D:** Doublet (Daraxonrasib + Mitazalimab) **Arm E:** Control (nal-IRI + 5-FU chemotherapy)	**Total:** 100,000 virtual patients per simulation run, conducted in triplicate. **Per Arm:** 20,000 patients.	**7 Predefined Archetypes:** ARCH-01: Young_Fit_Metastatic ARCH-02: Elderly_Frail_Metastatic ARCH-03: LAPC_Standard_Fitness ARCH-04: Young_Fit_BRCAm ARCH-05: Metastatic_KRAS_G12C ARCH-06: Metastatic_High_Stroma ARCH-07: Advanced_Refractory_PS1

Source: Synthesized from trial reports S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30.

Table 02: 3 Virtual Trials - Technical Specifications

C1: Drug Combination(s)	C2: Patient Data Granularity	C3: Modeling Architecture	C4: Project Timeline	C5: Primary Endpoints	C6: Key AI Models Utilized

R1: Details	**Core Triplet:** Daraxonrasib (KRAS G12C inhibitor) + Mitazalimab (immunotherapy) + liposomal Irinotecan. **Doublets & Control:** Various combinations of the core agents and a standard chemotherapy control were tested across the 5 arms.	Virtual patients were generated with a rich set of features defined by seven archetypes. Key data points included: age, disease stage (metastatic vs. locally advanced), ECOG performance status (0, 1, 2), tumor genomics (KRAS mutation status, specifically G12C; germline BRCA mutation status), and baseline tumor markers (CA 19-9).	An exponential survival model (Weibull shape k=1.0) was used to simulate time-to-event outcomes. Baseline hazards for the control arm were set to achieve median PFS of 3.1 months and OS of 6.1 months. Multiplicative hazard ratios (HRs) for each drug and a synergy factor (0.90) for the triplet were applied to model treatment effects.	The virtual trial simulations and analyses were conducted with a report date of July-August 2025. A fixed random seed (20250624) was used across all three trials to ensure reproducibility of the simulation runs.	**Co-primary Endpoints:** 1. **Overall Survival (OS):** Time from randomization to death from any cause. 2. **Progression-Free Survival (PFS):** Time from randomization to disease progression or death. **Secondary Endpoints:** 12-month OS rates and incidence of Grade ≥3 adverse events.	**For Cross-Verification & Meta-Verification:** 1. **grk4:** Grok 4 2. **grk3:** Grok 3 3. **ops4:** Opus 4 4. **g25p:** Gemini 2.5 Pro 5. **o3pr:** ChatGPT o3-pro

Source: Synthesized from trial reports S33, S37, S40 and verification files S43-S56.

Table 03: Benefits and Drawbacks of the 100K Patient Triplicate Simulation

C1: Itemized Benefits	C2: Itemized Drawbacks & Limitations

R1: Speed & Scale	**Unprecedented Scale:** Enabled the analysis of 100,000 patients per run, a scale infeasible for traditional trials, providing highly stable estimates for medians and HRs. **Rapid Execution:** The entire triplicate simulation and verification process was completed in a fraction of the time required for a physical trial, accelerating hypothesis testing.	**Model Simplification:** The exponential survival model is a simplification of complex real-world patient trajectories. It did not account for factors like dose modifications, treatment discontinuations due to AEs, or competing risks (as noted in S33).
R2: Robustness & Verification	**Triplicate Runs:** Running the simulation three times established the stability and low variance of the simulation engine. Cross-trial consistency scores were consistently high (S43-S47). **Multi-Model AI Verification:** The use of five diverse AI models (grk4, grk3, ops4, g25p, o3pr) to cross-verify all data provided an exceptionally robust, multi-perspective validation of the results and the verification process itself (S48, S55).	**Patient Profile Mismatch:** External validation against Flatiron data revealed a significant mismatch in baseline ECOG distribution, with the simulation having a healthier patient profile (fewer ECOG 2 patients). This limits the direct generalizability of the results to a real-world, less-fit population (S35b, S38b, S41b).
R3: Hypothesis Testing & Subgroup Analysis	**Complex Design Testing:** Allowed for the simultaneous evaluation of five complex regimens (a triplet and three doublets vs. control), which would be logistically challenging and costly in a traditional trial. **Deep Subgroup Analysis:** The large sample size enabled statistically robust analysis of seven distinct patient archetypes, confirming treatment benefits across subgroups and identifying a key treatment-genotype interaction (ARCH-05 / KRAS G12C).	**Reporting & Data Discrepancies:** The verification process consistently identified a major discrepancy between the reported KRAS-mutant (%) in the CSRs (~91%) and the kras_g12c=1 flag in the logs (~5%), indicating a need for clearer variable definitions and data capture protocols (S35, S38, S41).

R4: Future Improvements	**Data-Driven Trial Design:** The findings provide a quantitative framework to inform the design of future real-world trials, helping to prioritize the most promising regimens (e.g., Arm A vs. Arm D) and patient selection strategies.	**Refinement Opportunities:** Future simulations should incorporate more sophisticated models (e.g., agent-based models), include dynamic responses to toxicity, model biomarker efficacy more granularly (e.g., KRAS G12C benefit only applying to that subgroup), and refine the patient generation process to better match real-world ECOG distributions.

Source: Analysis of all provided files, with specific drawbacks cited.

Table 04: Reproducibility and Validation Findings

	C1: Validation (External Concordance)	**C2: Reproducibility (Internal & Cross-Model Consistency)**
R1: Overall Survival (OS)	**High Concordance:** The control arm (Arm E) from all three simulations demonstrated high external validity against Flatiron real-world data. OS% at all measured time points (0-24 months) fell within the ±5% pre-specified concordance threshold. The mean OS% difference was ~1.5% and the Pearson correlation was 0.999, both passing validation criteria (S35b, S38b, S41b).	**High Reproducibility:** Median OS values were extremely stable across the triplicate runs (e.g., Arm A mean OS of 8.73 mo, with a range of only 0.1 mo). Cross-trial consistency scores for Median OS and OS HR were high, averaging 8.98 and 9.08 respectively across the five AI models (S43-S47). This indicates the OS outcomes were highly reproducible.

R2: Baseline Characteristics	**Partial Concordance:** The simulated ECOG performance status distribution failed external validation. The absolute differences for ECOG 0, 1, and 2 vs. Flatiron data were ~5%, ~14%, and ~19% respectively, all exceeding the ±5% failure threshold (S35b, S38b, S41b). This indicates the simulated patient population was fitter than the real-world cohort.	**Exceptional Reproducibility:** Baseline characteristics were nearly identical across the three trials. Cross-trial consistency scores for all baseline metrics were ≥9.8 out of 10 across all AI models (S43-S47). The meta-verification analysis of the verification logs confirmed that discrepancies found were also highly consistent; for example, the KRAS-mutant deviation was found with a consistency score of 10.0 (S50-S54).
R3: Cross-Model Verification & Analysis	Not Applicable. External validation was performed on the simulation output itself, not on the AI models' analysis.	**Strong Inter-Model Agreement:** The five AI models showed remarkable agreement in their analyses. Visualizations confirmed a "tight cluster" for grk4, g25p, and o3pr, with grk3 and ops4 as minor outliers (S55). Agreement was highest for baseline metrics and lowest for archetype-specific outcomes (S48). The analysis included programmatic generation of visualizations (e.g., 01_heatmap_consistency_scores.py from S48) to quantify this agreement.
R4: Overall Reproducibility Assessment	The simulation's survival dynamics are externally valid, but the patient profile has limitations.	**Highly Robust:** The triplicate runs were highly consistent, with minimal variance in all primary and secondary endpoints. The AI-driven cross-trial verification process confirmed this stability with high consistency scores. Furthermore, meta-verification of the verification logs themselves also scored highly (mean scores >8.8), confirming the entire data generation and analysis pipeline is robust and reproducible (S50-S56). Analysis of visualization scripts (S49, S56) showed that percentage-based metrics (like AE rates) had higher consistency than time-to-event metrics (like median OS).

Source: Synthesis of all verification (S35, S38, S41), external validation (S35b, S38b, S41b), cross-trial verification (S43-S47, S48, S49), and meta-verification (S50-S54, S55, S56) files.

Key Insights

1. **Clinical Strategy:** The simulations provide strong evidence that a multi-pronged therapeutic strategy combining targeted therapy, immunotherapy, and chemotherapy (Arm A) yields the most significant survival benefit in advanced PDAC. The OS improvement of ~2.6 months over control is clinically meaningful. However, the associated 94% incidence of severe AEs makes this a high-risk, high-reward strategy. The chemo-free doublet of Daraxonrasib + Mitazalimab (Arm D) emerges as a highly promising alternative, offering a substantial survival benefit (OS HR ~0.76) with a more manageable toxicity profile, making it a potentially superior option for less-fit patients.
2. **Biomarker Importance:** The archetype sub-analyses powerfully validated the importance of biomarker-driven therapy. Patients in the KRAS G12C subgroup (ARCH-05) derived a clear and substantial benefit only from regimens containing the targeted inhibitor Daraxonrasib. This finding underscores the necessity of molecular profiling in PDAC to match patients with the most effective treatments and avoid ineffective therapies.
3. **Methodological Robustness:** This study exemplifies a new standard in virtual trial validation. By running the simulation in triplicate, performing internal log verification, validating against external real-world data, and employing a panel of five diverse AI models for cross-verification, the project established a high degree of confidence in the results. The process was robust enough to not only confirm the stability of the outputs but also to consistently identify its own limitations, such as the discrepancy in the KRAS definition and the mismatch in the ECOG profile.
4. **AI in Clinical Analysis:** The multi-model AI verification demonstrated both the power and the nuances of using AI for clinical data analysis. While there was remarkable consensus on objective metrics, minor systematic differences were observed between models, particularly in the calculation of variance-based consistency scores (S55, S56). This highlights the importance of using a multi-model approach or standardized consensus methods for critical analyses, mitigating the risk of relying on a single "black-box" algorithm. The success of this verification-of-verification process pioneers a new level of rigor for in-silico evidence generation.

Generation 57 (Plain Text)
Gemini 2.5 Pro
1969 words, 1 minutes 18.95 seconds

Virtual Study Overview: Analysis of 100,000-Patient Triplicate Simulations

Executive Summary

This report provides a comprehensive overview of a completed virtual study involving triplicate simulations of a 100,000-patient, 5-arm, Phase III-equivalent clinical trial in advanced Pancreatic Ductal Adenocarcinoma (PDAC). The study evaluated a novel triplet therapy (Daraxonrasib + Mitazalimab + liposomal Irinotecan) against various doublet

combinations and a standard-of-care control. Each of the three simulation runs was independently verified against its source logs and validated against external real-world data. The entire process was cross-verified by a suite of five independent AI models (grk4, grk3, ops4, g25p, o3pr) to ensure maximum robustness and reproducibility.

The clinical findings from the simulation consistently demonstrated that the triplet therapy (Arm A) conferred the greatest efficacy, with a mean Overall Survival (OS) of approximately 8.7 months versus 6.1 months for the control arm (OS Hazard Ratio [HR] ~0.69). This significant survival benefit was, however, accompanied by the highest rate of severe (Grade ≥3) adverse events (94% vs. 76.5% in control), highlighting a critical efficacy-toxicity trade-off.

Methodologically, the study establishes a new benchmark for in-silico trial validation. The triplicate runs were found to be exceptionally reproducible, with cross-trial consistency scores for key metrics consistently rated above 8.5 out of 10 by all AI models (S43-S47). External validation of the control arm against Flatiron data showed high concordance for survival outcomes (Pearson r = 0.999) but revealed a mismatch in baseline ECOG performance status, indicating an area for future model refinement (S35b, S38b, S41b). The multi-AI model verification process itself was highly reliable, confirming the stability of the simulation and analysis pipeline and underscoring the value of this approach for generating robust, defensible virtual evidence.

Technical Details

Table 01: 3 Virtual Trials - Overview

C1: Study Title/Identifier	C2: Primary Goal	C3: Trial Phase Equivalence	C4: Study Design	C5: Trial Arms	C6: Patient Population Size	C7: Patient Archetypes
R1: Details	A Phase III Virtual Study of Triplet Daraxonrasib + Mitazalimab + liposomal Irinotecan vs Doublets vs Chemotherapy in Advanced Pancreatic Ductal Adenocarcinoma (PDAC-SIM-001)	To compare the efficacy and safety of a novel triplet therapy against doublet combinations and standard chemotherapy control in advanced PDAC.	Phase: III (Virtual Simulation) Design: Randomized, controlled, parallel-group, five-arm study. Endpoints: Co-primary endpoints of Overall Survival (OS) and Progression-Free Survival (PFS) with a 24-month data cutoff.	5-arm in-silico simulation based on predefined patient archetypes and time-to-event models. Patients were randomized 1:1:1:1:1. Arm A: Triplet (Daraxonrasib + Mitazalimab + liposomal Irinotecan) Arm B: Doublet (Mitazalimab + liposomal Irinotecan) Arm C: Doublet (Daraxonrasib + liposomal Irinotecan) Arm D: Doublet (Daraxonrasib + Mitazalimab) Arm E: Control (nal-IRI + 5-FU chemotherapy)	Total: 100,000 virtual patients per simulation run, conducted in triplicate. Per Arm: 20,000 patients.7 Predefined Archetypes: ARCH-01: Young_Fit_Metastatic ARCH-02: Elderly_Frail_Metastatic ARCH-03: LAPC_Standard_Fitness ARCH-04: Young_Fit_BRCAm ARCH-05: Metastatic_KRAS_G12C ARCH-06: Metastatic_High_Stroma ARCH-07: Advanced_Refractory_PS1	

Source: Synthesized from trial reports S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30.

Table 02: 3 Virtual Trials - Technical Specifications

	C1: Drug Combination(s)	C2: Patient Data Granularity	C3: Modeling Architecture	C4: Project Timeline	C5: Primary Endpoints	C6: Key AI Models Utilized
R1: Details	Core Triplet: Daraxonrasib (KRAS G12C inhibitor) + Mitazalimab (immunotherapy) + liposomal Irinotecan. Doublets & Control: Various combinations of the core agents and a standard chemotherapy control were tested across the 5 arms.	Virtual patients were generated with a rich set of features defined by seven archetypes. Key data points included: age, disease stage (metastatic vs. locally advanced), ECOG performance status (0, 1, 2), tumor genomics (KRAS mutation status, specifically G12C; germline BRCA mutation status), and baseline tumor markers (CA 19-9).	An exponential survival model (Weibull shape k=1.0) was used to simulate time-to-event outcomes. Baseline hazards for the control arm were set to achieve median PFS of 3.1 months and OS of 6.1 months. Multiplicative hazard ratios (HRs) for each drug and a synergy factor (0.90) for the triplet were applied to model treatment effects.	The virtual trial simulations and analyses were conducted with a report date of July-August 2025. A fixed random seed (20250624) was used across all three trials to ensure reproducibility of the simulation runs.	Co-primary Endpoints: 1. Overall Survival (OS): Time from randomization to death from any cause. 2. Progression-Free Survival (PFS): Time from randomization to disease progression or death. Secondary Endpoints: 12-month OS rates and incidence of Grade ≥3 adverse events.	For Cross-Verification & Meta-Verification: 1. grk4: Grok 4 2. grk3: Grok 3 3. ops4: Opus 4 4. g25p: Gemini 2.5 Pro 5. o3pr: ChatGPT o3-pro

Source: Synthesized from trial reports S33, S37, S40 and verification files S43-S56.

Table 03: Benefits and Drawbacks of the 100K Patient Triplicate Simulation

	C1: Itemized Benefits	C2: Itemized Drawbacks & Limitations
R1: Speed & Scale	Unprecedented Scale: Enabled the analysis of 100,000 patients per run, a scale infeasible for traditional trials, providing highly stable estimates for medians and HRs. Rapid Execution: The entire triplicate simulation and verification process was completed in a fraction of the time required for a physical trial, accelerating hypothesis testing.	Model Simplification: The exponential survival model is a simplification of complex real-world patient trajectories. It did not account for factors like dose modifications, treatment discontinuations due to AEs, or competing risks (as noted in S33).
R2: Robustness & Verification	Triplicate Runs: Running the simulation three times established the stability and low variance of the simulation engine. Cross-trial consistency scores were consistently high (S43-S47). Multi-Model AI Verification: The use of five diverse AI models (grk4, grk3, ops4, g25p, o3pr) to cross-verify all data provided an exceptionally robust, multi-perspective validation of the results and the verification process itself (S48, S55).	Patient Profile Mismatch: External validation against Flatiron data revealed a significant mismatch in baseline ECOG distribution, with the simulation having a healthier patient profile (fewer ECOG 2 patients). This limits the direct generalizability of the results to a real-world, less-fit population (S35b, S38b, S41b).

R3: Hypothesis Testing & Subgroup Analysis Complex Design Testing: Allowed for the simultaneous evaluation of five complex regimens (a triplet and three doublets vs. control), which would be logistically challenging and costly in a traditional trial. Deep Subgroup Analysis: The large sample size enabled statistically robust analysis of seven distinct patient archetypes, confirming treatment benefits across subgroups and identifying a key treatment-genotype interaction (ARCH-05 / KRAS G12C). Reporting & Data Discrepancies: The verification process consistently identified a major discrepancy between the reported KRAS-mutant (%) in the CSRs (~91%) and the kras_g12c=1 flag in the logs (~5%), indicating a need for clearer variable definitions and data capture protocols (S35, S38, S41).

R4: Future Improvements Data-Driven Trial Design: The findings provide a quantitative framework to inform the design of future real-world trials, helping to prioritize the most promising regimens (e.g., Arm A vs. Arm D) and patient selection strategies. Refinement Opportunities: Future simulations should incorporate more sophisticated models (e.g., agent-based models), include dynamic responses to toxicity, model biomarker efficacy more granularly (e.g., KRAS G12C benefit only applying to that subgroup), and refine the patient generation process to better match real-world ECOG distributions.

Source: Analysis of all provided files, with specific drawbacks cited.

Table 04: Reproducibility and Validation Findings

C1: Validation (External Concordance) C2: Reproducibility (Internal & Cross-Model Consistency)

R1: Overall Survival (OS) High Concordance: The control arm (Arm E) from all three simulations demonstrated high external validity against Flatiron real-world data. OS% at all measured time points (0-24 months) fell within the ±5% pre-specified concordance threshold. The mean OS% difference was ~1.5% and the Pearson correlation was 0.999, both passing validation criteria (S35b, S38b, S41b). High Reproducibility: Median OS values were extremely stable across the triplicate runs (e.g., Arm A mean OS of 8.73 mo, with a range of only 0.1 mo). Cross-trial consistency scores for Median OS and OS HR were high, averaging 8.98 and 9.08 respectively across the five AI models (S43-S47). This indicates the OS outcomes were highly reproducible.

R2: Baseline Characteristics Partial Concordance: The simulated ECOG performance status distribution failed external validation. The absolute differences for ECOG 0, 1, and 2 vs. Flatiron data were ~5%, ~14%, and ~19% respectively, all exceeding the ±5% failure threshold (S35b, S38b, S41b). This indicates the simulated patient population was fitter than the real-world cohort. Exceptional Reproducibility: Baseline characteristics were nearly identical across the three trials. Cross-trial consistency scores for all baseline metrics were ≥9.8 out of 10 across all AI models (S43-S47). The meta-verification analysis of the verification logs confirmed that discrepancies found were also highly consistent; for example, the KRAS-mutant deviation was found with a consistency score of 10.0 (S50-S54).

R3: Cross-Model Verification & Analysis Not Applicable. External validation was performed on the simulation output itself, not on the AI models' analysis. Strong Inter-Model Agreement: The five AI models showed remarkable agreement in their analyses. Visualizations confirmed a "tight cluster" for grk4, g25p, and o3pr, with grk3 and ops4 as minor outliers (S55). Agreement was highest for baseline metrics and lowest for archetype-specific outcomes (S48). The

analysis included programmatic generation of visualizations (e.g., 01_heatmap_consistency_scores.py from S48) to quantify this agreement.

R4: Overall Reproducibility Assessment The simulation's survival dynamics are externally valid, but the patient profile has limitations. Highly Robust: The triplicate runs were highly consistent, with minimal variance in all primary and secondary endpoints. The AI-driven cross-trial verification process confirmed this stability with high consistency scores. Furthermore, meta-verification of the verification logs themselves also scored highly (mean scores >8.8), confirming the entire data generation and analysis pipeline is robust and reproducible (S50-S56). Analysis of visualization scripts (S49, S56) showed that percentage-based metrics (like AE rates) had higher consistency than time-to-event metrics (like median OS).

Source: Synthesis of all verification (S35, S38, S41), external validation (S35b, S38b, S41b), cross-trial verification (S43-S47, S48, S49), and meta-verification (S50-S54, S55, S56) files.

Key Insights

Clinical Strategy: The simulations provide strong evidence that a multi-pronged therapeutic strategy combining targeted therapy, immunotherapy, and chemotherapy (Arm A) yields the most significant survival benefit in advanced PDAC. The OS improvement of ~2.6 months over control is clinically meaningful. However, the associated 94% incidence of severe AEs makes this a high-risk, high-reward strategy. The chemo-free doublet of Daraxonrasib + Mitazalimab (Arm D) emerges as a highly promising alternative, offering a substantial survival benefit (OS HR ~0.76) with a more manageable toxicity profile, making it a potentially superior option for less-fit patients.

Biomarker Importance: The archetype sub-analyses powerfully validated the importance of biomarker-driven therapy. Patients in the KRAS G12C subgroup (ARCH-05) derived a clear and substantial benefit only from regimens containing the targeted inhibitor Daraxonrasib. This finding underscores the necessity of molecular profiling in PDAC to match patients with the most effective treatments and avoid ineffective therapies.

Methodological Robustness: This study exemplifies a new standard in virtual trial validation. By running the simulation in triplicate, performing internal log verification, validating against external real-world data, and employing a panel of five diverse AI models for cross-verification, the project established a high degree of confidence in the results. The process was robust enough to not only confirm the stability of the outputs but also to consistently identify its own limitations, such as the discrepancy in the KRAS definition and the mismatch in the ECOG profile.

AI in Clinical Analysis: The multi-model AI verification demonstrated both the power and the nuances of using AI for clinical data analysis. While there was remarkable consensus on objective metrics, minor systematic differences were observed between models, particularly in the calculation of variance-based consistency scores (S55, S56). This highlights the importance of using a multi-model approach or standardized consensus methods for critical analyses, mitigating the risk of relying on a single "black-box" algorithm. The success of this verification-of-verification process pioneers a new level of rigor for in-silico evidence generation.

Research Summary Generation 57: 728 Words

Outlining the Scope

I'm currently breaking down the request to create a comprehensive overview of the virtual study. My focus is on thoroughly understanding the scope of the project, which includes simulations with 100,000 patients, conducted in triplicate. I'm focusing on deconstructing the request and setting a solid foundation for the subsequent stages.

Developing a Strategy

I've crafted a comprehensive plan for extracting information from the provided documents. The initial focus is dissecting the request, determining the report structure, and extracting data from trial reports to construct the required markdown tables. Now, I'm setting a solid extraction strategy, working on each document to make sure I pull out the right information.

Structuring the Report

I'm now focusing on organizing the information into the required report format. My initial plan involves compiling a composite study title, goals, and design from the trial reports for the "Executive Summary" section. I'll summarize the key findings from each of the three trials, incorporating data from the visualizations. I'm prioritizing the creation of a clear and cohesive narrative from the complex data set.

Summarizing the Findings

I'm now fully immersed in synthesizing the data for the final report. I'm focusing on creating Table 04, which is pulling together findings from External Validations and Cross-Trial Verifications. I'm taking the OS concordance from external validations, and the consistency of outcomes from the Cross-Trial verifications. I'm also preparing to summarize key study insights, ensuring clear communication of complex results. This involves navigating and combining detailed information from a variety of sources to provide the most comprehensive overview of the virtual study.

Developing a Detailed Plan

I'm now fully immersed in outlining the scope, data extraction and overall structure. I'm focusing on deconstructing the comprehensive request and outlining a robust, multi-stage strategy. I plan to extract information for each markdown table, referencing the provided files and specified model abbreviations. My plan will incorporate all document types to ensure a thorough synthesis of the virtual study.

Synthesizing the Data

I'm now fully immersed in the synthesis process, focusing on integrating the complex data to craft a comprehensive virtual study overview. I've begun constructing Table 04, gathering data from external validations and cross-trial verifications. I'm focusing on the OS concordance from external data and outcome consistency from AI model verifications. The visualizations will support the report.

Developing the Tables

I'm now deep in the process of building the required markdown tables. I've begun to populate Table 01 using information from trial reports, summarizing study titles, goals, and design for a clear overview. I'm focusing on the precise data extraction from all three trials, ensuring accuracy.

Refining the Approach

I'm currently engaged in refining my information extraction strategy, aiming for a more robust method. I've broken down the diverse file types: trial reports, log verifications, and cross-trial visualizations, for a structured extraction process. My next step will be to prioritize the information flow, focusing on what should go into the executive summary, technical details, and key insights sections.

Organizing the Information

I'm currently focused on extracting the crucial information for Table 04, which involves integrating data from external validations and cross-trial verifications. I'm prioritizing the synthesis of external OS concordance metrics and the consistency scores from the AI models. This will allow for a comprehensive report.

Refining the Synthesis

I'm now fully immersed in synthesizing the data for Table 04, which focuses on reproducibility and validation. I'm prioritizing the integration of insights from the meta-verification process, particularly the consistent findings related to discrepancies across the trials. I'm aiming to synthesize the data from diverse files.

Focusing on Table Development

I've been heavily engaged in the construction of Table 04, which is central to summarizing the study's reproducibility and validation. I'm focusing on external OS concordance and the consistency metrics from the AI cross-verifications. I am integrating insights from the meta-verification process to bolster my analysis. I am creating a clear and accurate report of the complex data.

Structuring the Report

I've made significant progress in defining the report structure and content, focusing on creating the four required markdown tables. I'm prioritizing the compilation of a concise "Executive Summary" and the structuring of a robust "Technical Details" section, which will contain the detailed tables summarizing study design, findings, benefits, drawbacks, reproducibility, and validation data. I've also drafted the key insights for the "Key Insights" section, ensuring a clear, insightful narrative.